7 Deadly Hadoop Misconfigurations

Kathleen Ting | February 2013

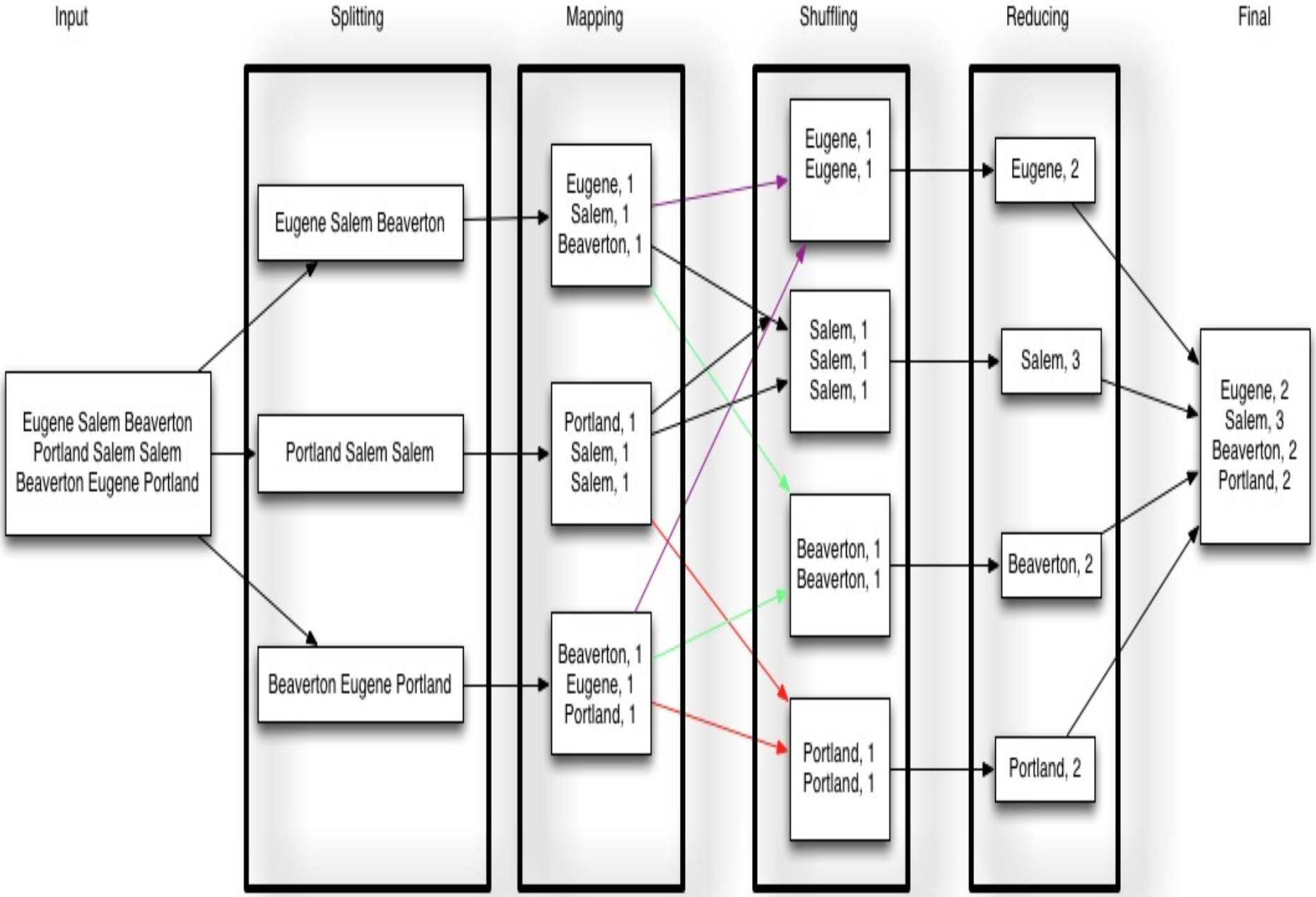
Who Am I?

Kathleen Ting

Apache Sqoop Committer, PMC Member

Customer Operations Engineering Mgr, Cloudera

@kate_ting, kathleen@apache.org


Input
Splitting
Mapping
Shuffling
Reducing
Final
Eugene Salem Beaverton Portland Salem Salem Beaverton Eugene Portland
Eugene Salem Beaverton
Portland Salem Salem
Beaverton Eugene Portland
Eugene, 1 Salem, 1 Beaverton, 1
Portland, 1 Salem, 1 Salem, 1
Beaverton, 1 Eugene, 1 Portland, 1
Eugene, 1 Eugene, 1
Salem, 1 Salem, 1 Salem, 1
Beaverton, 1 Beaverton, 1
Portland, 1 Portland, 1
Eugene, 2
Salem, 3
Beaverton, 2
Portland, 2
Eugene, 2 Salem, 3 Beaverton, 2 Portland, 2

Agenda

- Ticket Breakdown
- What are Misconfigurations?

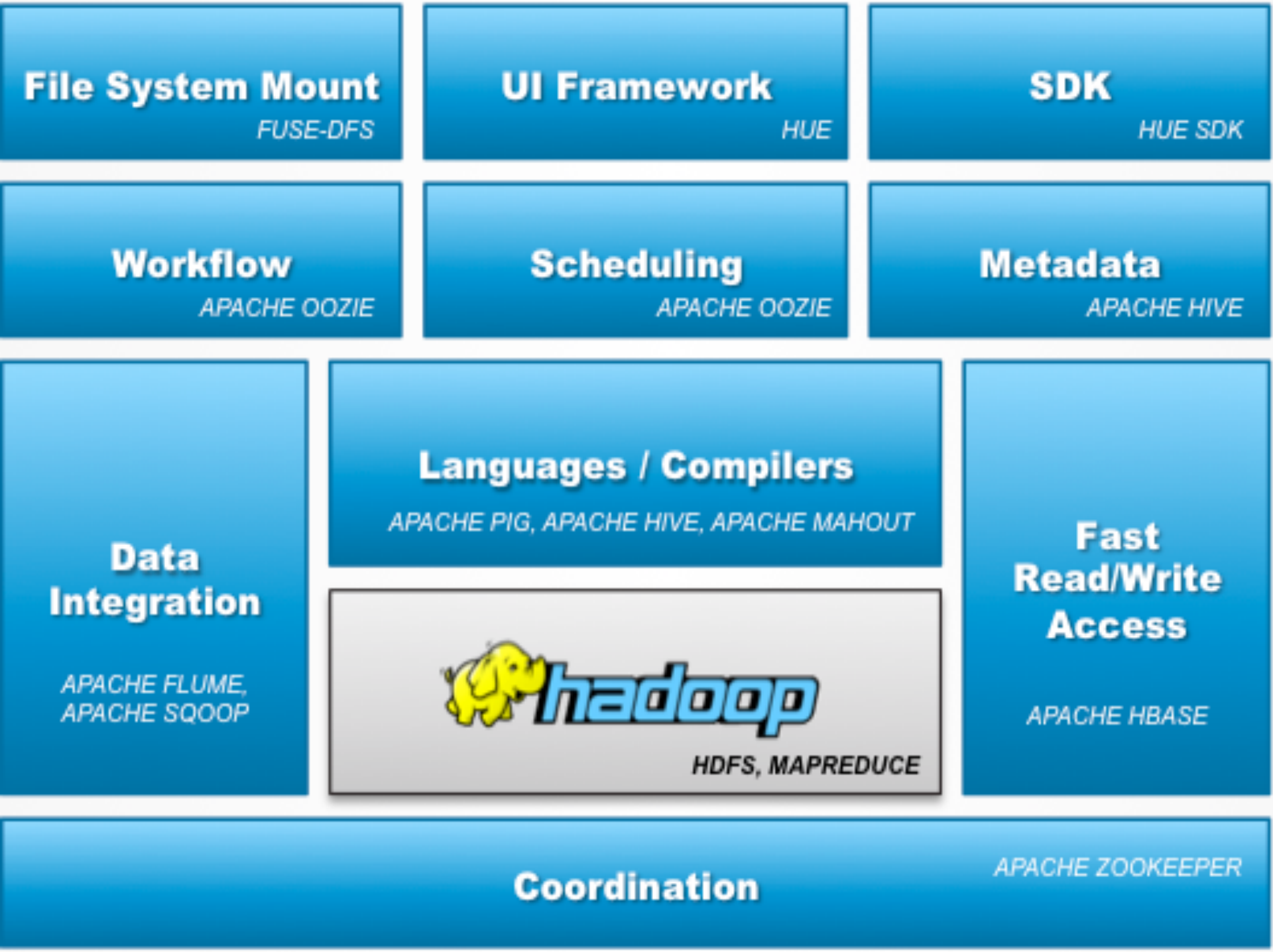
  Memory Mismanagement
  - TT OOME
  - JT OOME
  - Native Threads

  Thread Mismanagement
  - Fetch Failures
  - Replicas

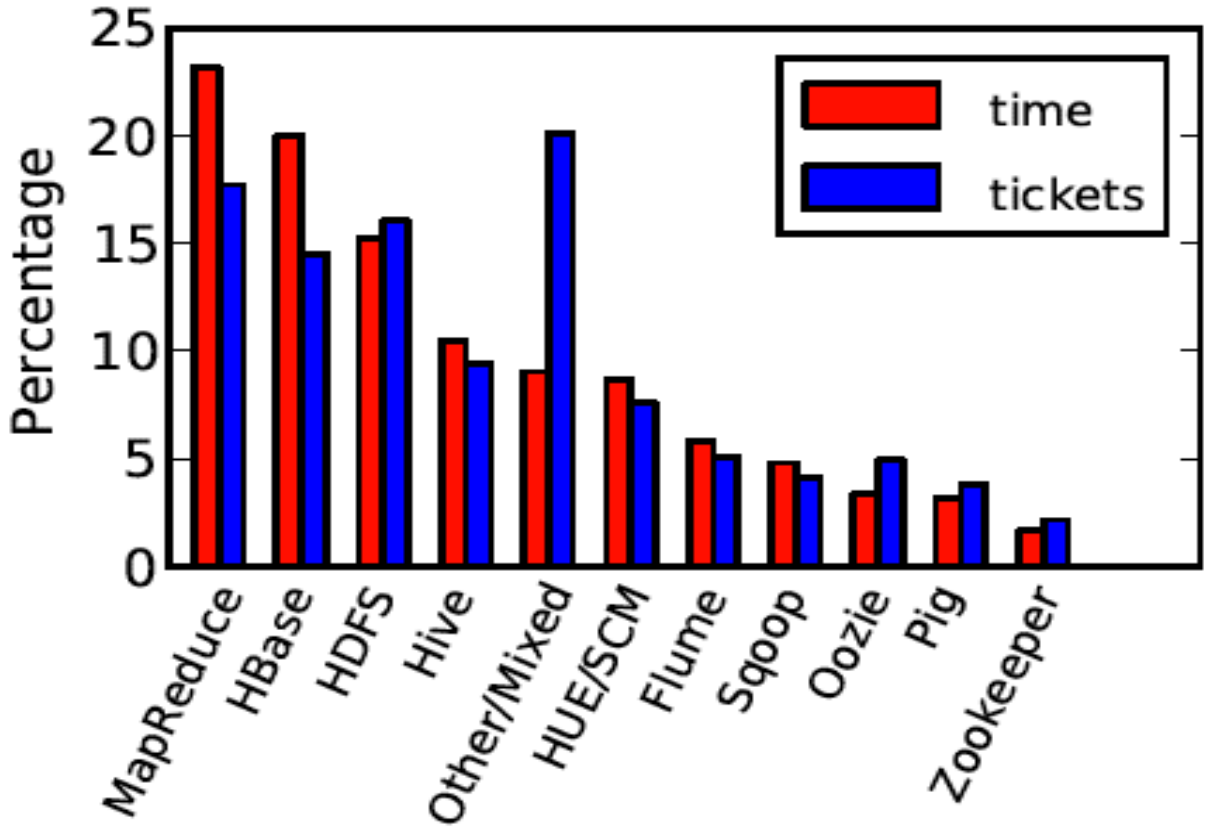
  Disk Mismanagement
  - No File
  - User Error

Agenda

- Ticket Breakdown
- What are Misconfigurations?
  - Memory Mismanagement
    - TT OOME
    - JT OOME
    - Native Threads
  - Thread Mismanagement
    - Fetch Failures
    - Replicas
  - Disk Mismanagement
    - No File
    - User Error


File System Mount
FUSE-DFS
UI Framework
HUE
SDK
HUE SDK
Workflow
APACHE OOZIE
Scheduling
APACHE OOZIE
Metadata
APACHE HIVE
Data Integration
APACHE FLUME, APACHE SQOOP
Languages / Compilers
APACHE PIG, APACHE HIVE, APACHE MAHOUT
hadoop
HDFS, MAPREDUCE
Fast Read/Write Access
APACHE HBASE
Coordination
APACHE ZOOKEEPER

By Tickets Filed, MapReduce is Central to Hadoop

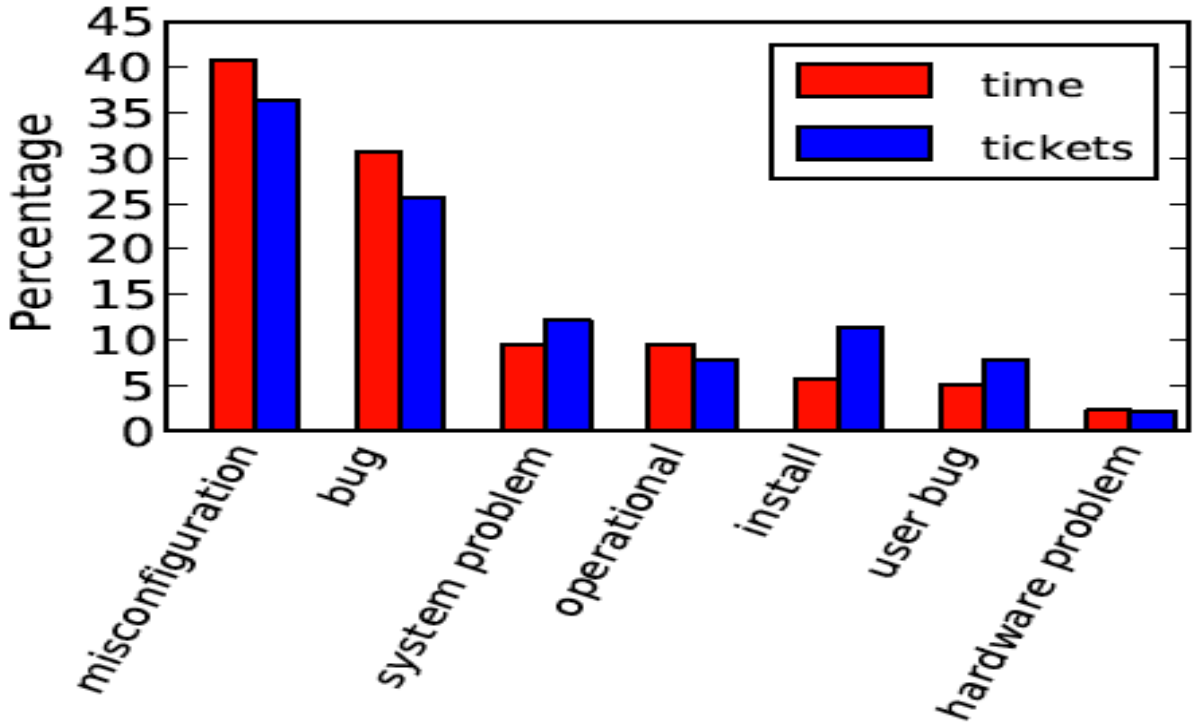


Agenda

- Ticket Breakdown
- What are Misconfigurations?
  - Memory Mismanagement
    - TT OOME
    - JT OOME
    - Native Threads
  - Thread Mismanagement
    - Fetch Failures
    - Replicas
  - Disk Mismanagement
    - No File
    - User Error

What are Misconfigurations?

Issues requiring change to Hadoop or to OS config files
Comprises 35% of Cloudera Support Tickets
e.g. resource-allocation: memory, file-handles, disk-space



Why Care About Misconfigurations?

The life of an over-subscribed MR/Hive cluster is nasty, brutish, and short.

(with apologies to Thomas Hobbes)

What else you got?


```
FAILED: Execution Error, return
code 2 from
org.apache.hadoop.hive.ql.exec.MapR
edTask
```

Faulty MR Config Killed Hive

Shuffle phase for query failed.

Heap increased but not buffer size.

They had `io.sort.mb = 112M`

Should be `io.sort.mb = 512M`

Agenda

- Ticket Breakdown
- What are Misconfigurations?

  Memory Mismanagement
  - TT OOME
  - JT OOME
  - Native Threads

  Thread Mismanagement
  - Fetch Failures
  - Replicas

  Disk Mismanagement
  - No File
  - User Error

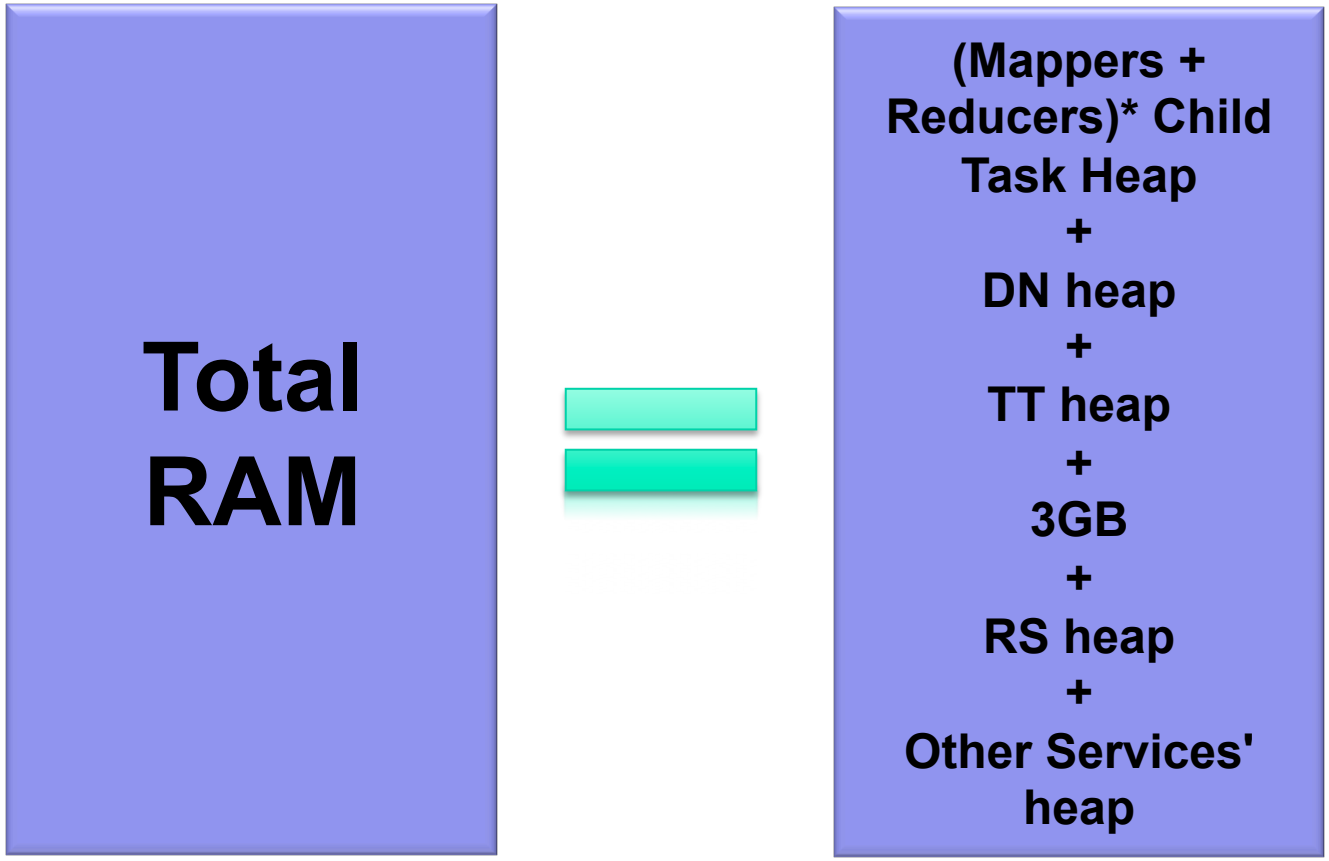
1. Task Out Of Memory Error


```
FATAL org.apache.hadoop.mapred.TaskTracker:
Error running child : java.lang.OutOfMemoryError:
Java heap space
       at org.apache.hadoop.mapred.MapTask
$MapOutputBuffer.<init>
```

- What does it mean?
  - Memory leak in task code
- What causes this?
  - MR task heap sizes will not fit

1. Task Out Of Memory Error

- TaskTracker side
  - mapred.child.ulimit > 2*mapred.child.java.opts
  - 0.25*mapred.child.java.opts < io.sort.mb < 0.5*mapred.child.java.opts
- DataNode side
  - Use short pathnames for dfs.data.dir names
    - e.g. /data/1, /data/2, /data/3
  - Increase DN heap


Total RAM
=
(Mappers + Reducers)* Child Task Heap
+
DN heap
+
TT heap
+
3GB
+
RS heap
+
Other Services' heap

2. JobTracker Out of Memory Error

```
ERROR org.apache.hadoop.mapred.JobTracker: Job
initialization failed:
java.lang.OutOfMemoryError: Java heap space
at
org.apache.hadoop.mapred.TaskInProgress.<init>(TaskInProg
ress.java:122)
```

- What does it mean?
  - Total JT memory usage > allocated RAM
- What causes this?
  - Tasks too small
  - Too much job history

2. JobTracker Out of Memory Error

- How can it be resolved?
  - sudo -u mapreduce jmap -histo:live <pid>
  - Increase JT heap
  - Don’t co-locate JT and NN
  - mapred.job.tracker.handler.count = ln(#TT)*20
  - mapred.jobtracker.completeuserjobs.maximum = 5
  - mapred.job.tracker.retiredjobs.cache.size = 100
  - mapred.jobtracker.retirejob.interval = 3600000

3. Native Threads


```
ERROR mapred.JvmManager: Caught Throwable in JVMRunner.
Aborting TaskTracker.
java.lang.OutOfMemoryError: unable to create new native thread


ERROR org.apache.hadoop.hdfs.server.datanode.DataNode:
java.io.IOException: Too many open files
```

- What does it mean?
  - DN show up as dead even though processes are still running on those machines
- How can it be resolved?
  - In /etc/security/limits.conf adjust low settings for open files, process, or max memory
  - Recommend setting is 64k+

Agenda

- Ticket Breakdown
- What are Misconfigurations?

Memory Mismanagement

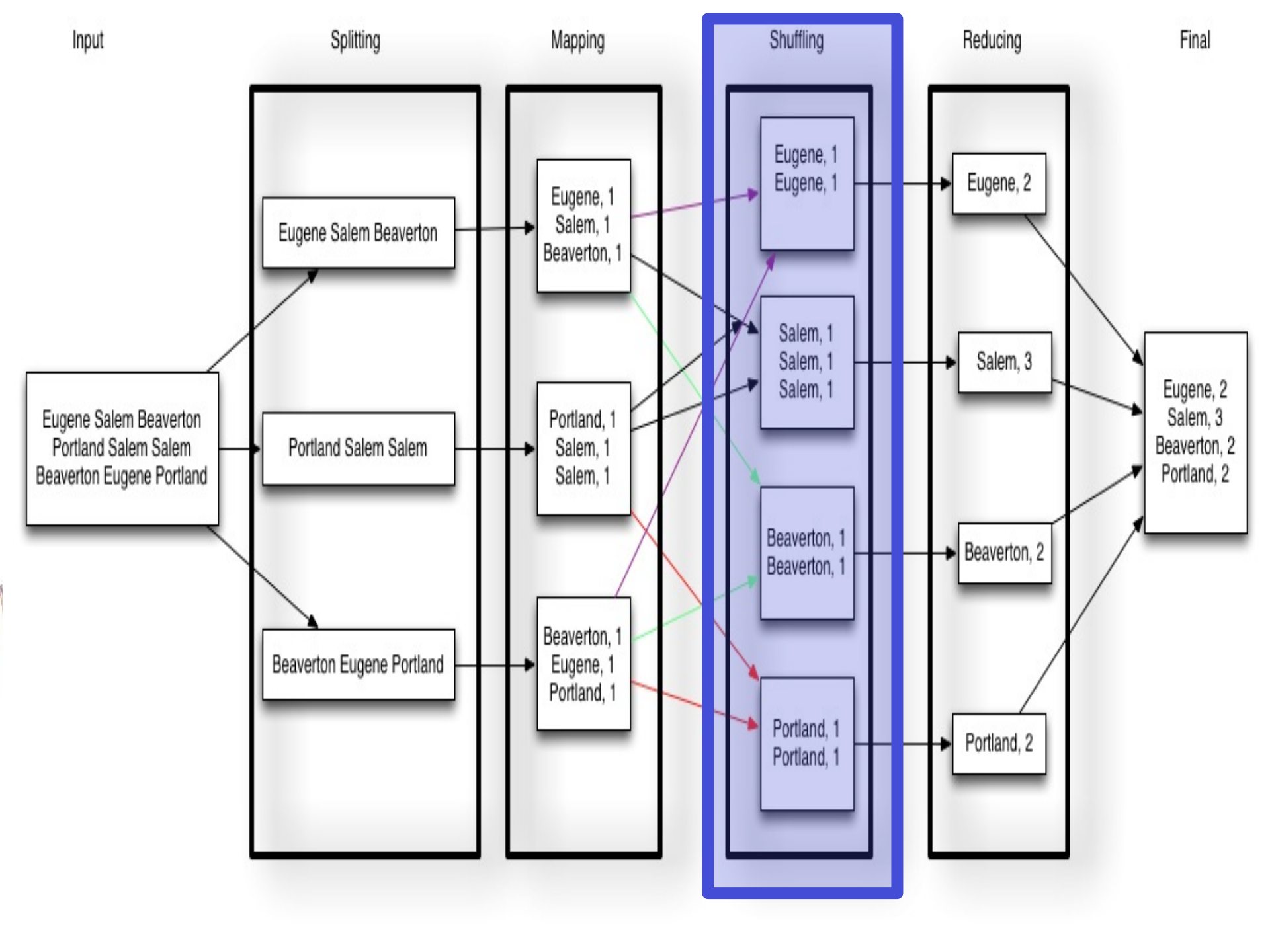
- TT OOME
- JT OOME
- Native Threads

Thread Mismanagment

- Fetch Failures
- Replicas

Disk Mismanagement

- No File
- User Error


Input
Splitting
Mapping
Shuffling
Reducing
Final
Eugene Salem Beaverton Portland Salem Salem Beaverton Eugene Portland
Eugene Salem Beaverton
Portland Salem Salem
Beaverton Eugene Portland
Eugene, 1 Salem, 1 Beaverton, 1
Portland, 1 Salem, 1 Salem, 1
Beaverton, 1 Eugene, 1 Portland, 1
Eugene, 1 Eugene, 1
Salem, 1 Salem, 1 Salem, 1
Beaverton, 1 Beaverton, 1
Portland, 1 Portland, 1
Eugene, 2
Salem, 3
Beaverton, 2
Portland, 2
Eugene, 2 Salem, 3 Beaverton, 2 Portland, 2

4. Too Many Fetch-Failures


```
INFO org.apache.hadoop.mapred.JobInProgress:
Too many fetch-failures for output of task
```

- What does it mean?
  - Reducer fetch operations fail to retrieve mapper outputs
  - Too many could blacklist the TT
- What causes this?
  - DNS issues
  - Not enough http threads on the mapper side
  - JVM bug

4. Too Many Fetch-Failures

- How can it be resolved?
  - mapred.reduce.slowstart.completed.maps = 0.80
  - tasktracker.http.threads = 80
  - mapred.reduce.parallel.copies = SQRT(Nodes), floor of 10
  - mapred.tasktracker.shuffle.fadvise = false (CDH3u3)
  - Stop using 6.1.26 Jetty

5. Not Able to Place Enough Replicas

```
WARN
org.apache.hadoop.hdfs.server.namenode.FSNamesystem
: Not able to place enough replicas
```

- What causes this?
  - dfs replication > # available DNs
  - Block placement policy
  - DN being decommissioned
  - Not enough xcievers threads

5. Not Able to Place Enough Replicas

How can it be resolved?

- dfs.datanode.max.xcievers = 4096
- Look for nodes down (or rack down)
- Check disk space
- Rebalance under-replicated blocks
  - dfs.namenode.replication.work.multiplier.per.iteration = 30
  - dfs.balance.bandwidthPerSec = 10MB/s
  - Move files from full volume to empty volume

Agenda

- Ticket Breakdown
- What are Misconfigurations?

Memory Mismanagement

- TT OOME
- JT OOME
- Native Threads

Thread Mismanagement

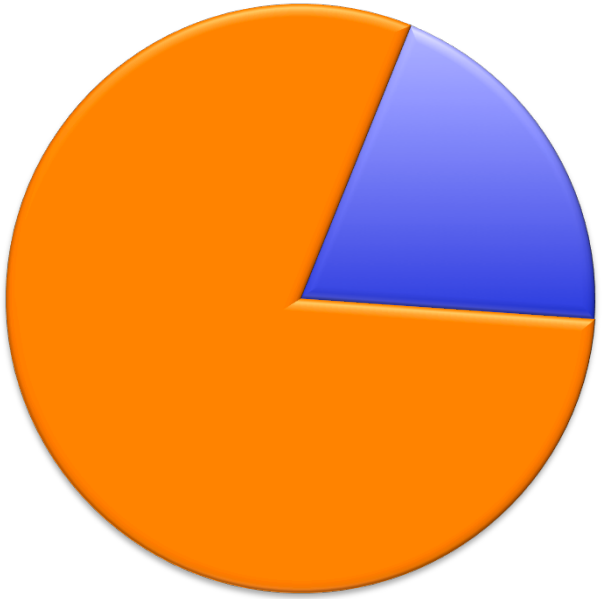
- Fetch Failures
- Replicas

Data Mismanagement

- No File
- User Error

6. No Such File or Directory


```
ERROR org.apache.hadoop.mapred.TaskTracker: Can not start task
tracker because ENOENT: No such file or directory
at org.apache.hadoop.io.nativeio.NativeIO.chmod(Native Method)
```



6. No Such File or Directory

What does it mean?

TT failing to start or jobs are failing

What causes this?

TT filling

Wrong permissions

Bad disk

How can it be resolved?

dfs.datanode.du.reserved = 10%

Permissions = 755, owner = mapred

7. User Error

Accidentally issued: `hadoop fs -rmr /data/`

Permanent data loss unless fs.trash.interval configured

- Default of 0 = permanent loss
- Set to 1440 min so contents stick around for a day
- Reference: HDFS-3302, HDFS-2740, HADOOP-8598

Bonus: Dr. Who


```
WARN
org.apache.hadoop.security.UserGroupInformat
ion: No groups available for user dr.who
```

ACLs required for viewing job details
Unauthenticated user = "dr. who”

How can it be resolved?

- Pass specific user via URL
- Configure Kerberos
- (Tweak hadoop.http.staticuser.user from dr.who default)

Takeaways

Correct configuration is up to you.

Misconfigurations are hard to diagnose.

Get it right the first time with monitoring tools.

"Yep - we were able to download/install/ configure/setup a Cloudera Manager cluster from scratch in minutes :)"